웹 크롤링을 사용한 자동화된 이미지 분류 모델

이주혁*, 김미희*
*한경대학교 컴퓨터응용수학부
{xpdlwm99, mhkim}@hknu.ac.kr

Automated Image Classification Model Using Web Crawling

Ju-Hyeok Lee*, Mi-Hui Kim*
*School of Computer Engineering & Applied Mathematics, Hankyong National University

요 약

최근 딥러닝은 이미지 인식, 음성 인식 등 여러 분야에서 고려되고 있는 기술이다. 그러나 딥러닝 기술을 이용하기 위해서는 대형데이터 세트가 필요하나 이를 구축하기 힘들고 많은 시간이 필요하다는 문제점이 있다. 이에, 본 논문에서는 웹 크롤링을 통해 사용자가 원하는 카테고리의 이미지 데이터 세트를 수집하고 수집한 데이터들을 전처리 과정을 통해 딥러닝 모델에 입력할 수 있는 데이터 세트의 구축을 자동화하며, 전이학습을 통해서 적은 훈련 시간과 높은 정확도를 얻을 수 있는 이미지 분류모델을 제안한다.

1. 서론

딥러닝은 사람의 신경세포처럼 기계가 학습할 수 있도록 하는 인공신경망을 이용한 기계학습 방법이며, 이미지 인식, 음성 인식 등 여러 분야에서 사용중이다[1]. 대부분 제안된 이미지 인식 방법들에서는 합성곱 신경망(이하 “CNN”이라고 한다.)을 사용한다. CNN은 인공신경망을 이용하여 이미지 내에서 추출한 특성들을 이용해 결과를 매칭하는 시스템을 가진다[2]. 딥러닝을 학습시키기 위한 데이터는 대형 데이터 세트가 필요하지만 이러한 대형데이터 세트를 자신이 원하는 이미지의 데이터 세트를 구하기도 힘들며, 직접 데이터 세트를 만들기 위해서는 많은 시간이 필요하다[3].

본 논문에서는 이러한 데이터 구축 문제를 해결하기 위해서 웹 크롤링을 이용해서 자신이 정한 카테고리의 이미지 데이터 세트 구축을 자동화하고 카테고리의 데이터 세트를 전이학습을 통해서 적은 훈련 시간과 정확도를 높이는 이미지 분류모델을 제안한다. 또한, 실험을 통해 CNN과 비교하여 성능 향상을 보인다.

2. 배경 지식

2.1 웹 크롤링

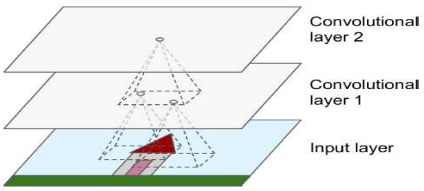
온라인 상에는 대용량의 비정형 데이터를 포함한 다양한 데이터가 쌓여있다. 이런 데이터를 수집하고 분석 및 활용하는 기술인 빅 데이터도 발전을 하고 있다[4]. 빅 데이터에서 가장 중요한 것은 데이터이며, 이 데이터를 수집하는 방법 중에는 크롤링이라는 기술이 있다. 크롤링이란 무수히 많은 컴퓨터나 온라인 상에 나뉘어져 있는 문서나 이미지 등 검색의 대상이 되는 데이터들을 모으는 기술이다. 크롤링 중에서도 웹상에서 데이터를 추출하는 것을 웹 크롤링이라고 하며[5], 이를 정형화 또는 자동화한 시스템을 크롤러라고 한다.

본 논문에서 제안하는 웹 크롤링은 웹상에서 사용자가 원하는 카테고리의 데이터를 추출하는 기술을 자동화한 웹 크롤러이다.

2.2 CNN

CNN은 대뇌의 시각 피질 연구에서 시작되었다가 1980년대부터 이미지 인식 분야에서 사용됐다. CNN에서 가장 중요한 구성 요소는 합성곱 층(Convolutional layer)이다. 이 합성곱 층에서 입력 이미지의 픽셀들이 합성곱 층 뉴런의 수용장 안에 픽셀에만 연결되어서 특성에 집중하게 되며, 이렇게 추출된 작은 특성들에서 층을 점점 거치면서 고수준 특성들로 조합해 결국 고수준 특성들을 이용해 CNN이 인식률이 높게 측정된다[6]. CNN에는 합성

곱 층 이외에도 필터, 커널, 패딩, 풀링 층 등 여러 층이 모여서 하나의 CNN을 형성한다. 자세한 CNN의 구성 요소에 대한 설명은 3.2절 이미지 분류모델에서 (그림4)에서 어떤 층이 사용되었는지 확인하겠다. (그림1)은 합성곱 층의 예시이다.



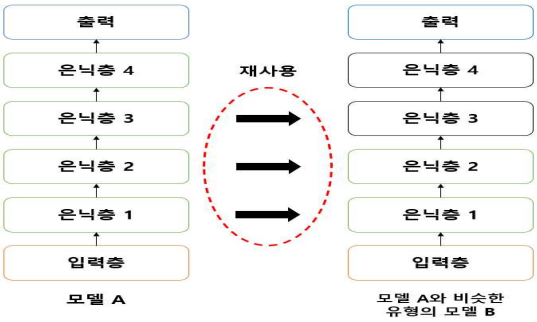
(그림 1) CNN 내의 합성곱 층 예시[6]

본 논문에서는 이 CNN을 이미지 분류 모델의 전이학습에 사용이 되며, 전이학습의 정확도와 훈련 시간을 실험 부분에서 비교한다.

2.3 전이학습

여러 딥러닝 모델들이 제작이 되면서, 모델 간 작업이나 목적이 비슷한 경우가 생긴다. 비슷한 유형의 문제를 해결하는 모델에서 신경망의 하위층을 재사용하여, 처음부터 새로 훈련하는 수고를 덜어줄 수 있는데 이를 전이학습이라고 한다. 전이학습은 작업이 비슷할수록 더 많은 층을 재사용할 수 있으며, 아주 비슷한 작업이라면 모든 은닉층을 유지하고 출력층만 교체가 가능한 장점이 있다[6]. (그림 2)는 모델 A와 비슷한 유형의 모델을 가진 B가 모델 A의 사전 훈련된 층을 재사용할 때의 예시이다.

본 논문에서는 이미지를 분류하는 카테고리 즉 클래스의 이름만 다르며 하나의 클래스를 분류하는 모델이기에 전이학습을 통해서 새로운 모델을 만들 수 있다. 작업이 유사할수록 많은 층을 재사용할 수 있고 출력층만 교체할 수 있기에 CNN 모델을 처음부터 새로운 데이터로 학습할 수고를 줄일 수 있다. 전이학습을 통해 하나의 클래스의 특징을 추출하는 층을 재사용하여 처음부터 새로 훈련하는 수고를 줄이고, 분류 정확도는 높이기 위해 전이학습을 사용한다.



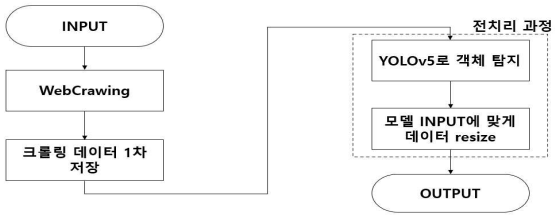
(그림 2) 사전훈련된 층 재사용하기

3. 제안하는 모델

3.1 웹 크롤러 모델

제안하는 모델 중 웹 크롤러에 대한 모델은 (그림 3)과 같다. 웹 크롤러의 입력 값으로는 사용자가 원하는 카테고리의 이름을 영문으로 입력받는다. 입력받은 카테고리를 기반으로 웹 크롤링을 시작한다. 웹 크롤링은 selenium[7]과 chromedirver[8]를 통해서 구글의 이미지 검색 창으로 넘어가 해당 카테고리를 검색하고, 그에 해당하는 모든 자료를 수집한다. 수집한 자료들은 1차 저장하여 데이터 전처리 과정을 거친다. 데이터 전처리를 하는 이유는 웹 크롤링 특성상 비정형화된 데이터가 많으며, 해당 입력값의 내용을 검색했지만, 입력값의 내용과는 전혀 이미지가 검색됐을 수도 있어 전처리 과정을 거친다.

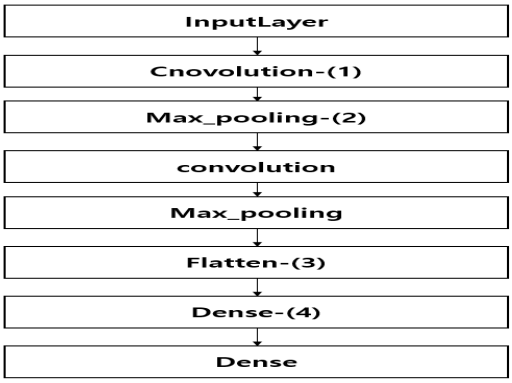
전처리 과정에는 YOLOv5[9]를 사용한다. YOLOv5 모델에는 여러 종류의 카테고리 이미지가 학습 한다. YOLOv5를 이용해 이미지 내의 객체를 탐지한 결과가 탐지되지 않거나, 카테고리가 2개 이상 탐지되는 경우는 이상치의 데이터로 판단하여 제거했다. 카테고리가 2개 이상 탐지되는 경우도 제거한 이유는 CNN의 합성곱 층에서 특성을 추출할 때, 방해될 수 있기에 제거한다. 전처리 과정을 거치게 되면 CNN을 이용한 전이학습의 모델 입력 값에 맞게 픽셀의 크기를 조절해주면 사용자 카테고리의 데이터가 완성된다.



(그림 3) 웹 크롤러 흐름도

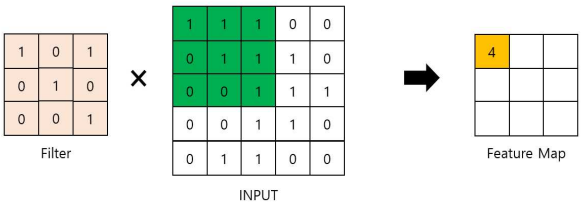
3.2 이미지 분류 모델

먼저 CNN을 통해 이미지 분류 모델의 틀을 설계한다. 틀을 먼저 설계하는 이유는 전이학습을 사용한다고 하더라도, 층을 재사용하기 위해서는 원본 모델이 있어야 층을 전달할 수 있기에 CNN 모델을 먼저 설계한다. (그림4)는 본 논문에서 이미지 분류 모델로 제안하는 CNN의 모델 구조이다.



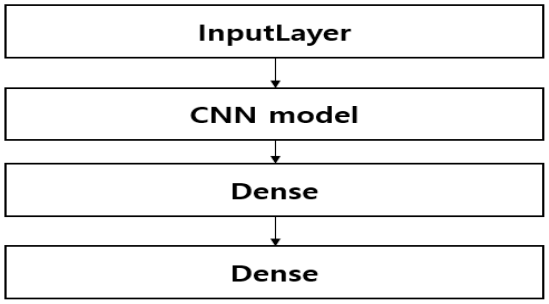
(그림 4) CNN 모델 구조

CNN모델에서 사용된 계층은 합성곱 층, 맥스풀링 층, 플래튼 층, 은닉층을 거쳐 마지막으로 출력층으로 구성되어 있다. 합성곱 층(1)은 (그림1)처럼 사용이 되며, (그림 5)처럼 합성곱 연산을 통해서 이미지의 특징들을 추출한다. 그 다음 맥스 풀링층(2)을 거치게 되는데, 맥스 풀링층에서는 합성곱 층의 출력 데이터를 입력으로 받는다. 이 출력 데이터의 크기를 줄이거나 특징들을 강조하기 위해 사용하며, 특히 맥스풀링 층은 입력받은 데이터에서 특정 영역 안에 최댓값을 모아 특징들이 뚜렷해지게 해주며 공간 방향 차원을 줄이는 역할을 한다. 플래튼 층(3)은 이미지를 1차원으로 줄여주는 역할을 한다. 합성곱 층이나 맥스풀링 층을 반복하면 주요 특징들을 추출할 수 있는데, 추출된 주요 특징을 은닉층으로 전달해줘야 하는 역할로 바로 플래튼층이 사용된다. 플래튼 층을 통해서 은닉층(4)으로 전달하게 되고 은닉층과 활성화 함수를 통해 마지막으로 이미지를 분류하는 모델을 가지게 된다.



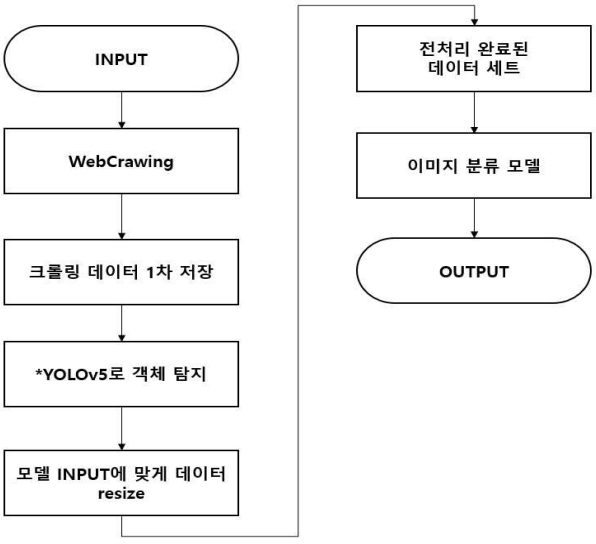
(그림 5) 합성곱 연산 예시

위에서 설계한 CNN 모델을 바탕으로 학습 시간을 줄이고 정확도를 높이기 위한 모델을 설계를 위해 전이학습 사용하여 (그림 6)처럼 (그림 4)의 CNN 모델의 층이 삽입된 전이학습이 완료된 모델의 모습이다. (그림 4)에서 합성곱 층과 맥스풀링 층을 거치고 플래톤 층을 거쳐 특징들을 전달해주는 계층들을 모아서 특징들을 고정하고 전이시켰다.



(그림 6) 전이학습을 거친 이미지 분류 모델

(그림 7)은 본 논문에서 제안하는 모델의 흐름도이다. 입력 값으로 사용자가 원하는 카테고리를 입력하면 전처리 완료된 데이터 세트까지는 (그림 3)의 웹 크롤러의 과정을 처리한 후에 (그림 6) 모델의 입력 값 데이터로 사용이 되며 최종 결과로는 분류 결과가 나오게 된다.



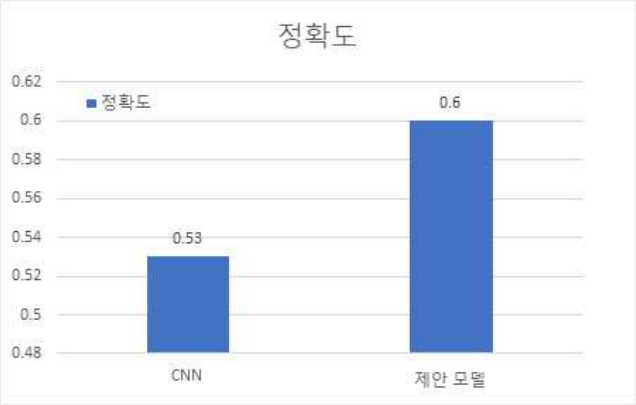
(그림 7) 제안하는 모델의 흐름도

4. 실험 결과

카테고리를 예시로 강아지로 정하고 웹 크롤러를 통해 전 처리된 총 12,335장의 이미지가 모였다. 실험 모델은 위에서 제작한 (그림 3) CNN과 (그림 6)의 이미지 분류 모델 2가지로 비교 실험을 진행하였

으며, 모델의 정확도와 훈련 시간을 비교했다.

(그림 8)에서 보이는 것처럼 정확도 결과로는 CNN이 0.53의 정확도를 제안 모델은 0.6의 정확도를 얻었다. 정확도가 높아진 이유는 CNN에 비해 특성에 대한 학습을 더 할 수 있으며, 카테고리에 맞는 미세조정에 의해서 정확도가 높아진다. (그림 9)에서 보이는 것처럼 훈련 시간의 결과로는 CNN은 35초의 시간이 제안 모델은 10초의 시간이 걸렸다. 훈련 속도가 빨라진 이유 또한 CNN처럼 처음부터 데이터를 학습하는 것이 아니기에 CNN보다 모델 훈련 속도가 빠르게 측정된다.



(그림 8) 정확도

(그림 9) 모델 훈련 시간

5. 결론 및 향후 연구

본 논문에서는 웹 크롤링을 사용한 자동화 이미지 분류모델을 제안하였다. 웹 크롤링 과정에서 YOLOv5에 속해있지 않은 데이터의 경우는 학습이 필요하며, 간혹 불필요한 데이터도 속해있는 경우가 있다.

전이학습으로 만들어진 모델을 학습할 때 모델 미세 조정을 사용자가 직접해야 하는데 이 부분도 자동화를 통해 적합한 미세 조정을 구현하고자 한다. 또한, 훈련 시간은 전이학습의 기대효과가 나타났지만, 정확도에서는 큰 효과가 나타나지 않았다. 문제점은 CNN 모델과 제안 모델의 구조가 너무 단조로운 점이 있다. 향후 연구에서는 전처리 과정과 CNN 구조를 보완하고, 복수 카테고리까지 이미지 분류 가능한 모델을 구현하고자 한다.

6. Acknowledgement

이 논문은 2018년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No.2018R1A2B6009620), 교신저자 김미희.

참고문헌

[1] 안신영, 박유미, 임은지, 최완, "딥러닝 분산처리 기술동향", [ETRI] 전자통신동향분석 31(3), 131-141, 2016.

[2] 허이룬, "합성곱 신경망(CNN)기반 이미지 처리 시스템," 국내박사학위논문 배재대학교, 2018.

[3] 권현, 김용철, "딥러닝 모델에 대한 적대적 사례 기술 동향," 정보보호학회지 31(2), 5-12, 2021.

[4] 김정숙, "빅 데이터 활용과 관련기술 고찰". 한국콘텐츠학회지, 10(1), 34-40, 2016.

[5] 서동민, 정한민, "빅데이터 분석 서비스 지원을 위한 지능형 웹크롤러," 한국콘텐츠학회논문지,13(12), 575-584, 2013.

[6] Aurelien Geron, "Hands-On machine Learning with Scikit-Learn, Keras & TensorFlow 2nd Edition," O'Reilly Media, 2019.

[7] Selenium, https://www.selenium.dev/ko/, 2021.

[8] Chromedriver, https://sites.google.com/a/chromium.org/chromedriver/, 2021.

[9] Joseph Redmon, Santosh Divvala, Ross Girshick, Ali Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 779-788, 2016.